Traitement automatique d'un corpus de récits de voyages pyrénéens : Analyse syntaxique, sémantique et pragmatique dans le cadre de la théorie des types

Anaïs Lefeuvre, Richard Moot, Christian Retoré

Traitement automatique d'un corpus de récits de voyages pyrénéens : Analyse syntaxique, sémantique et pragmatique dans le cadre de la théorie des types

Anaïs Lefeuvre, Richard Moot, & Christian Retoré, {prenom.nom@labri.fr}

Université de Bordeaux, LaBRI-CNRS & INRIA

Ce travail de recherche a reçu un soutien financier d' INRIA et du Conseil Régional d'Aquitaine dans le cadre du projet Itipy

1 Introduction

Cet article décrit les étapes qui composent notre analyse du discours, en partant du texte brut, et pour en produire une représentation sémantique dans le cadre de la Discourse Representation Theory, désormais DRT (Kamp and Reyle, 1993). Une chaîne complète de traitement est proposée et testée sur le corpus Itipy, «Itinéraires Pyrénéens», lequel a été proposé par la médiathèque de Pau. Le premier but applicatif consiste à attacher un lieu aux portions de texte narrant une action dans ce lieu. Nous exploitons alors ce corpus de récits de voyage du XIXème siècle dans l'objectif d'extraire automatiquement les itinéraires décrits et afin d'indexer les portions de texte prenant effectivement pour décors les lieux géographiques en question.

Notre outil, Grail est un parser pour grammaire logique de types avec un ensemble restreint de règles fixes et utilisant un lexique riche. Tout d'abord, la première phase a consisté en l'acquisition de la grammaire sur un corpus annoté (Paris 7 Treebank). Ce corpus nous a permis d'obtenir les informations grammaticales propres aux unités du lexique de la langue française présentes dans le corpus, le lexique produit ne contient donc pas la totalité des mots du français et contient plusieurs catégories pour les entrées les plus fréquentes. Dans la chaine de traitement, la méthode d'attribution de la catégorie intègre une approche statistique : lorsqu'un mot est absent du lexique, l'analyse propose une catégorie ou lorsqu'il présente plusieurs catégories possibles, elle sélectionne la plus appropriée.

Chaque mot du texte est taggé, puis supertaggé en fonction des autres unités se trouvant dans son contexte proche (la phrase). Le supertagger propose plusieurs formules qui correspondent à une analyse syntaxique partielle pour chaque phrase du texte dans le cadre des grammaires catégorielles, et plus précisément du calcul de Lambek. S'ensuit une étape de combinaison de toutes les analyses partielles pour donner l'analyse globale. La structure obtenant la meilleure probabilité étant sélectionnée, on garde cette structure comme organisation du calcul de la représentation sémantique en fonction des unités qui la composent. On associe alors à chaque mot son λ-terme à partir du lexique sémantique cette fois et dont la formule correspond à celle présente dans le lexique grammatical pour cette même entrée (Moot, 2010). Le λ-terme pour chaque unité sémantique est saisi à la main dans le style de la λ-DRT. La représentation sémantique étant produite automatiquement à partir de l'analyse syntaxique, nous obtenons une représentation logique sémantique bien formée.

La dimension pragmatique quant à elle ne peut être reléguée à un plan inférieur dans l'interprétation du discours. En effet, une analyse du discours impose de fait une interaction entre la sémantique des unités de langue dont on doit interpréter le sens en discours et la prise en compte de la dimension pragmatique de ce qui est dit. Notre approche s'inspire de l'approche de Busquets et al. (2001), "une théorie de l'interprétation des discours doit être aussi en fait une théorie de la sémantique, de la pragmatique, et de leur interaction, c'est-à-dire une théorie de l'interface pragmatique-sémantique" . Certains phénomènes sémantiques restent cependant difficiles à traiter, certains cas de glissement de sens montrent qu'une flexibilité dans le typage

doit être permise, alors que dans les cas les plus courants le typage doit être rigide pour éviter une représentation inappropriée. Nous donnerons quelques exemples à propos et proposons donc afin d'améliorer les résultats de notre chaîne traitement de traiter ces phénomènes par l'affinement des λ-termes du lexique dans le cadre du système F, λ-calcul d'ordre supérieur.

Nous détaillerons ici notre corpus et nos objectifs applicatifs quant à celui-ci, nous présenterons les étapes de traitement du discours, commençant par l'acquisition de la grammaire du français sur corpus annoté, puis l'analyse syntaxique dans le cadre des grammaires catégorielles. Nous expliquerons plus amplement l'interface syntaxe-sémantique dans la théorie des types logiques permettant la construction de nos représentations sémantiques en λ-DRT. Nous présenterons le système F et notre traitement des phénomènes discursifs mettant en jeu l'interaction sémantique-pragmatique puis nous présenterons les perspectives de ce travail.

2 Le corpus

Notre corpus de 576 334 mots est une collection de 11 oeuvres classées par la médiathèque de Pau comme récits de voyages pyrénéens du XIXème et début XXème siècle. Présentons rapidement les données textuelles de notre corpus : le genre du récit de voyage, implique de fait une hétérogénéité interne reconnue, certains spécialistes désignent par ailleurs le récit de voyage comme un "genre fragmenté" (Magri-Mourgues, 2009), dans lequel on trouve une myriade de procédés narratifs (ici Pasquali (1994) reprend la théorie d'Affergan dans Exotisme et altérité, 1987) incluant "le récit métonymique" ,"le récit synecdochique", "le récit métaphorique", "le récit de voyage et de découverte du réel", etc. Ajoutons à ceci que le corpus Itipy est constitué de récits écrits par des géologues, des topographes (Vincent de Chausenque), ou encore des romanciers (George Sand). Malgré la diversité des formes de discours qui composent le corpus, sa spécificité réside dans le récit de l'itinéraire, seul point commun entre tous les textes. La structure narrative du récit de voyage observe une alternance entre la description de l'itinéraire emprunté et d'autres informations telles que des observations sur le relief, le caractère des personnages rencontrés ou encore des considérations introspectives du narrateur sur des domaines variés.

Notre traitement du discours se détache totalement des données du genres, ou encore de la nature des thèmes abordés. Nous intervenons alors sur l'analyse profonde syntaxique et sémantique, mais aussi pragmatique et discursive, des fragments de récits identifiés comme pertinents. Notre visée est applicative sur l'extraction des itinéraires et permet une mise à l'épreuve pour nos modèles logiques. Le lexique grammaticale est acquis automatiquement tandis qu'une partie du lexique sémantique est saisi à la main pour les objets complexes nécessitant un raffinement lexical.

3 L'acquisition de la grammaire du français et l'analyse syntaxique

Grail est une plateforme pour le développement et le parsing des grammaires catégorielles multimodales (Moot, 1998). Il a surtout été utilisé pour développer des grammaires spécifiques à des phénomènes linguistiques, souvent en exploitant la transparence de l'interface syntaxe/sémantique des grammaires catégorielles.

Il est nécessaire de préciser que la grammaire pour le français a été semi-automatiquement extraite à partir du corpus de Paris 7 Treebank (Abeillé et al., 2003). L'acquisition de cette grammaire a permis d'obtenir un lexique pour le français donnant une catégorie syntaxique dépendante de son contexte à chacun des mots du corpus. Le tagger et le supertagger ont été entrainés sur la grammaire extraite et atteignent une précision de 98,4% pour le tagger et de 90,5% pour le supertagger (Moot, 2010).

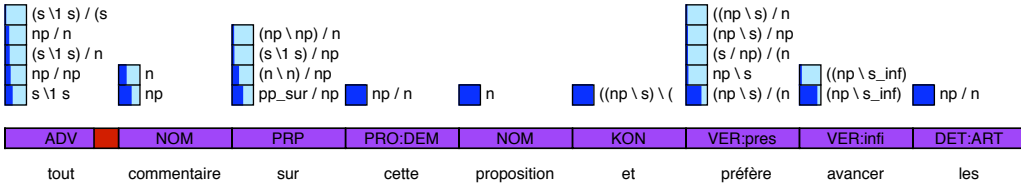
Le traitement démarre donc par le tagging, chaque mot se voit associer une catégorie syntaxique (part-of-speech), puis le supertagging permet de proposer pour chaque phrase plusieurs analyses syntaxiques en arbre. On notera que les items lexicaux les plus représentés ont souvent plusieurs formules associées, par exemple "et" ou "est". Dans le cas où le mot est inconnu, une catégorie est proposée par le supertagger

ce qui permet l'acquisition de nouveau items. Les catégories syntaxiques sont attribuées dans le cadre du calcul de Lambek, elles sont proposées en fonction des catégories environnantes afin d'obtenir la catégorie S pour la phrase. Le calcul est effectué selon les règles de la logique intuitionniste telles que décrites dans (Retoré, 2000). Un des atouts du calcul de Lambek réside dans le fait que le mot est considéré comme une unité contextuelle, dépendant du contexte et que par l'attribution même des catégories aux mots, on obtient la structure syntaxique de la phrase. Ce calcul nous permet de construire une analyse bien formée.

Pour illustrer le résultat du supertagging sur le discours, voici un exemple tiré du journal Le Point[1] :

(1) Le gouvernement refuse tout commentaire sur cette proposition et préfère avancer les chiffres positifs récoltés par la mesure.

Figure 1- Exemple de sortie du supertagger pour l'exemple présenté



La figure 1 montre une partie de la sortie du supertagger. Les probabilités de chaque supertag en fonction de leur contexte proche sont indiquées de manière non exhaustive dans les boîtes associées aux catégories, les plus probables étant les plus foncées.

Le nombre de supertags par mot reste raisonnable : 2,3 tags par mot pour les 98,4% de supertags corrects. Par exemple, "et" n'a qu'une seule catégorie associée, celui d'une conjonction entre VP prenant un VP ou $np\backslash s$ à sa droite pour créer un nouveau $np\backslash s$. Lorsque les catégories sont plus difficiles à discriminer pour un même item, on aura donc plusieurs formules associées, on pense ici aux adverbes ADV, les prépositions PRP et les différentes formes de verbes VER :X.

Pour la préposition "sur", 4 catégories sont proposées à Grail sur les 16 formules associées à cet item : le premier choix du supertagger étant pp_{sur}/np, la catégorie pour une préposition qui sélectionne un syntagme nominal à sa droite pour créer un pp argument du verbe, et son second choix, $(n\backslash n)/np$, la catégorie pour une préposition modifiant un nom. Les deux autres possibilités ne sont pas choisies.

L'utilisation des probabilités permet de choisir la meilleure analyse globale. Dans l'ordre décroissant des probabilités pour chaque catégorie, le suppertagger teste la compatibilité de l'analyse avec ses catégories voisines jusqu'à résolution de la phrase. C'est exactement à partir de la structure syntaxique choisie que l'on va pouvoir associer une représentation sémantique du discours.

4 La dérivation sémantique en λ-DRT

Tout d'abord, la DRT est une théorie proposant de représenter la sémantique d'un discours grâce à un modèle présenté comme une boîte (Discourse Representation Structure) dans laquelle on trouve premièrement le domaine, composé des individus, puis les conditions d'interprétation sémantique de ce modèle. Les DRS créées se fusionnent les unes avec les autres par l'opération de "merge" et permettent d'interpréter les phénomènes de cohérence du discours comme par exemple la résolution des anaphores pronominales.

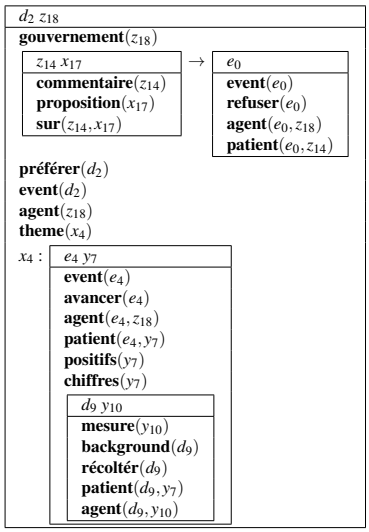
L'isomorphisme Curry-Howard (Howard, 1980) montre que les dérivations en grammaires catégorielles sont des sous-ensembles des dérivations de la logique intuitionniste. Comme introduit précédemment, les catégories syntaxiques dans le style du calcul de Lambek permettent donc d'associer une lecture sémantique exprimée par le λ-terme simplement typé dans le style de la DRT correspondant à chaque mot taggé.

La description en λ-DRT est écrite à la main et introduite dans le lexique pour les items lexicaux nécessitant un tel traitement. Pour chaque item au sein de la structure syntaxique, on associe à sa catégorie le terme en λ-DRT associé puis on réduit l'expression dans l'ordre de la structure.

En dérivant l'analyse sémantique de l'analyse syntaxique, on conserve la bonne formation de la représentation de l'expression correspondant exactement à la catégorie et on observe aussi le principe de compositionnalité (Montague, 1974).

En premier lieu, il convient de définir ce qu'est un type selon la sémantique de Montague. L'analyse syntaxique de type *S* (qui correspond au type sémantique des phrases, les valeurs de vérités **t**) est un λ-terme dont les variables libres correspondent aux mots, et le lexique fournit des λ-termes du même type sémantique : en les substituant et en réduisant le terme obtenu on obtient un terme normal de type **t**. C'est une formule logique, la représentation sémantique, et dans notre cas la λ-DRS. Néanmoins de manière général, il faut au minimum partager le type **e**, les individus (aussi appelés entités), en diverses sortes, pour que le calcul de la sémantique bloque à juste titre lorsque le type d'un argument ne correspond pas au type attendu par la fonction. Par exemple on peut demander que le type du sujet de *descend* soit un objet animé[2]. Regardons la représentation sémantique de l'exemple 1.

Figure 2- La DRS de l'exemple 1



Les variables d, e et f sont des variables d'évènement et les variables x, y et z sont des variables d'entités. Ici, est la représentation pour l'évènement "préférer" dont l'agent est z_{18}, lui même une variable dont la propriété associée est "gouvernement". Le thème de l'évènement d_2, c'est à dire la situation "préférée" est la DRS imbriquée x_4. On remarque que z_{18} est agent de cette DRS imbriquée et agent de la DRS emboitée à droite de l'implication puisque notre exemple contient deux syntagmes verbaux dont "le gouvernement" est sujet.

Le typage du lexique permet de vérifier de manière stricte la bonne formation de la représentation mais ne permet pas d'interpréter certaines expressions du discours parfois plus souples de ce point de vue, c'est pourquoi nous proposons une solution dans la suite des travaux de Pustejovsky (1995), Luo (2011) et d' Asher (2011) en sémantique lexicale.

5 Les phénomènes traités par le système F

Dans un travail initié par Bassac et al. (2010) dédié à la partie sémantique lexicale de l'analyse, un système complet d'organisation du lexique a été proposé afin de traiter le discours de manière aussi proche que possible des travaux de Montague (1974). Ces travaux ont été motivé par les phénomènes de coprédication suivant :

(2) Le dîner était sympathique, pourtant l'entrée était brûlée.

Ici, on réfère à deux aspects d'un événement complexe, le dîner.

(3) Ce livre est volumineux mais intéressant.

Coprédication correcte entre les deux facettes de livre : contenu informationnel et objet physique.

(4) J'ai mis les livres au grenier, je les avais tous lus.

Les livres sont comptés en tant qu'objets physiques, puis repris par *les* en tant que contenus informationnels par le second prédicat.

(5) Washington borde le Potomac et a attaqué l'Irak.

Coprédication incorrecte (sauf trait d'humour) sur le président siégeant à Washington et le lieu géographique de cette même ville.

Pour ces phénomènes, nous avons conçu une structure de lexique et un algorithme qui permettent de calculer les représentations sémantiques de telles phrases, de rendre compte des coprédications correctes (2, 3, 4) et d'échouer lorsqu'elles ne le sont pas (5), de quantifier correctement.

Pour le genre de phénomènes que nous souhaitons étudier, nous nous plaçons dans le cadre du λ-calcul du second ordre appelé système F (Girard, 1971). Ce formalisme nous permettra de manipuler plus finement les types, de quantifier sur eux, d'utiliser la coercion afin de résoudre les cas de coprédication par exemple.

Tout d'abord, nous devons détailler les types du second ordre tels que nous les utilisons :

- Sont des types de base :
  - $\mathbf{t}$ *les valeurs de vérité*, $\mathbf{v}$ les *événements*,
  - quelques autres types constants correspondant aux *différentes sortes d'individus*, par exemple *chemin*,
  - des *variables de type*, notées par des lettres grecques (issues d'un ensemble dénombrable P)
- Lorsque T est un type et α une variable de type, qui peut ou non apparaître dans T, $\Pi\alpha.\ T$ est un type (dit polymorphe).
- Lorsque T_1 et T_2 sont des types, $T_1 \to T_2$ est aussi un type.

Pour définir les termes, on se donne une infinité dénombrable de variables de chaque type, ainsi que des constantes en nombre fini pour chaque type :[3]

- Une variable de type T c'est-à-dire $x : T$ (ce qu'on écrit aussi x^T) est un *terme* de type T.
- Une constante de type T c'est-à-dire $c : T$ (ce qu'on écrit aussi c^T) est un *terme* de type T.
- $(f\ \tau)$ est un terme de type U quand τ est de type T et f de type $T \to U$.
- $\lambda x^T.\ \tau$ est un terme de type $T \to U$ quand x est une variable de type T, et τ un terme de type U.
- $\tau\{U\}$ est un terme de type $T[U/\alpha]$ quand $\tau : \Lambda\alpha.\ T$, et U est un type.
- $\Lambda\alpha.\tau$ est un terme de type $\Pi\alpha.T$ quand α est une variable de type $\tau : T$ sans occurrence de α dans le type d'une variable libre.

Lorsque les constantes sont celles de la logique multisorte d'ordre supérieur (opérateurs $\& : \mathbf{t} \to \mathbf{t} \to \mathbf{t}, \forall_i : (e_i \to \mathbf{t}) \to \mathbf{t}, \forall_{i,j} : (e_i \to (e_j \to \mathbf{t}) \to \mathbf{t}, \ldots)$ et constantes du langage logique *regarde* : $e_h \to e_h \to \mathbf{t}$) ce système est appelé $\blacksquare\mathsf{Ty}_{\mathbf{n}}$.

Les réductions pour λ et Λ sont définies de manières similaires :

- $(\lambda x.\tau)u$ se réduit en $\tau[u/x]$ (réduction habituelle).
- $(\Lambda\alpha.\tau)\{U\}$ se réduit en $\tau[U/\alpha]$ (rappelons que α est une variable de type et U un type quelconque).

On a les résultats bien connus suivants :

- Tout terme du système F admet une et une seule forme normale (Girard, 1971).
- Corollaire : si les constantes (du λ-calcul) correspondent au langage L multisorte d'une logique d'ordre n (opérations logiques, prédicats, fonctions et constantes), tout terme normal de type $\mathbf{t}$ correspond à une formule de L.

Que signifient-ces résultats ? L'analyse syntaxique d'une phrase $m_1 \cdots m_n$ étant de type S, sa contrepartie sémantique est un terme $u[x_1, \ldots, x_n]$ de type $\mathbf{t}$ dont les variables libres $x_i : X_i$ correspondent aux types des mots. En remplaçant dans u les $x_i : X_i$ par les termes principaux fournis par le lexique, on va obtenir un terme de type $\mathbf{t}^4$, dont on calcule la forme normale t° : les résultats ci-dessus nous garantisse que t° est bel et bien une formule logique du langage L : c'est la représentation sémantique de la phrase.

Le lexique contient un terme principal, qui s'apparente à celui de la sémantique de Montague, hormis qu'il est écrit dans $\Lambda\mathsf{Ty}_n$. Il contient aussi des termes optionnels, qu'on utilise au besoin en cas de conflit de types, ceux-ci correspondant aux glissements de sens évoqués dans les exemples. Une bonne partie de ces termes sont de simples fonctions qui transforment un objet de type x en un objet de type y lorsque x est un sous-type de y (par exemple g : voiture $\rightarrow$ véhicule).

Les représentations sémantiques de la phrase sont obtenues comme suit :

1. Obtenir une analyse sémantique qui soit au minimum un arbre précisant pour chaque nœud interne quel sous arbre s'applique à quel autre sous-arbre : ce que font les grammaires catégorielles comme expliqué dans la section précédente, La dérivation sémantique en λ-DRT.
2. L'écrire comme un λ-terme de $\Lambda\mathsf{Ty}_n$.
3. Résoudre les conflits de types en utilisant les λ-termes de $\Lambda\mathsf{Ty}_n$ comme indiqué ci-après, plusieurs solutions sont possibles, ce sont tous les termes correctement typés de type $\mathbf{t}$.
4. Réduire ces λ-termes, en vertu des résultats ci-dessus mentionnés des formules logiques, ce sont les représentations sémantiques de l'énoncé analysé.

Les conflits se présentent sous la forme $(\lambda x^A.u)w^W$: un terme de type A est attendu par la fonction $(\lambda x^A.u)$ mais l'argument fourni est de type W. Pour résoudre les conflits, on procède de l'une des deux manières suivantes :

La transformation rigide : le lexique fournit, pour un mot impliqué dans u ou pour un mot impliqué dans w un terme g de type $W \rightarrow A$: le terme se résout en $(\lambda x^A.u)(gw)^A$.

La transformation flexible : les diverses occurrences de x^A dans u sont utilisées avec des types différents $A_1, \ldots, A_n$: on peut utiliser, si le lexique en fournit, des termes différents de types $g_i : W \rightarrow A_i$ pour chaque occurrence de x et remplacer comme le veut la β-réduction chaque occurrence de x par $(g_i(w)) : A_i$.

Dans l'intérêt de la simplicité de l'exemple choisi, nous ne prenons pas en compte cette nuance dans le présent article, néanmoins le mécanisme est effectif, il a été implanté dans Grail par Emeric Kien (Kien, 2010).

5.1 Le voyageur fictif

Nous nous demandons par ailleurs dans quelle mesure ces phénomènes ont une incidence dans le traitement des itinéraires. Très fréquemment dans le corpus, les chemins ou les routes qui sont des objets statiques portent le déplacement, ce qui nous pousse à proposer un voyageur fictif sous-jacent :

(6) (...) cette route monte jusqu'à Lux où l'on arrive par une jolie avenue de peupliers.

(7) (...) cette route qui monte sans cesse pendant deux heures

(8) Le chemin descend pendant deux heures (...) *Non extrait du corpus, mais similaire, et plus rapide à traiter ci-après.*

Dans l'exemple 6 on pourrait penser que la phrase signifie simplement que l'altitude de la route est une fonction croissante de l'abscisse curviligne, du point de référence jusqu'à *Lux*. Mais le second exemple

7 montre clairement que cette interprétation n'est pas convenable, il faut obligatoirement considérer un voyageur qui suit cette route à cause du circonstant *pendant deux heures*. On observe dans le corpus que ce genre de constructions où la route devient le voyageur qui la suit, peut très bien s'appliquer alors que le narrateur, lui même en voyage ne suit pas la route ainsi décrite ! Il faut même parfois suivre assez longtemps la description avant de pouvoir dire si le voyageur la suit ou non. Le voyageur en question n'est donc pas forcément l'un des référents de discours : il peut être fictif et doit être introduit dans le terme et lié par une quantification (existentielle en hypothèse ou universelle en conclusion).

Comme dans les cas précédents, le modèle détecte la nécessité d'un glissement de sens par un conflit de types, dont la nature diffère des précédents :

$$\left(P^{person\rightarrow \mathbf{t}}\left(u^{path}\right)\right) \qquad person \neq path$$

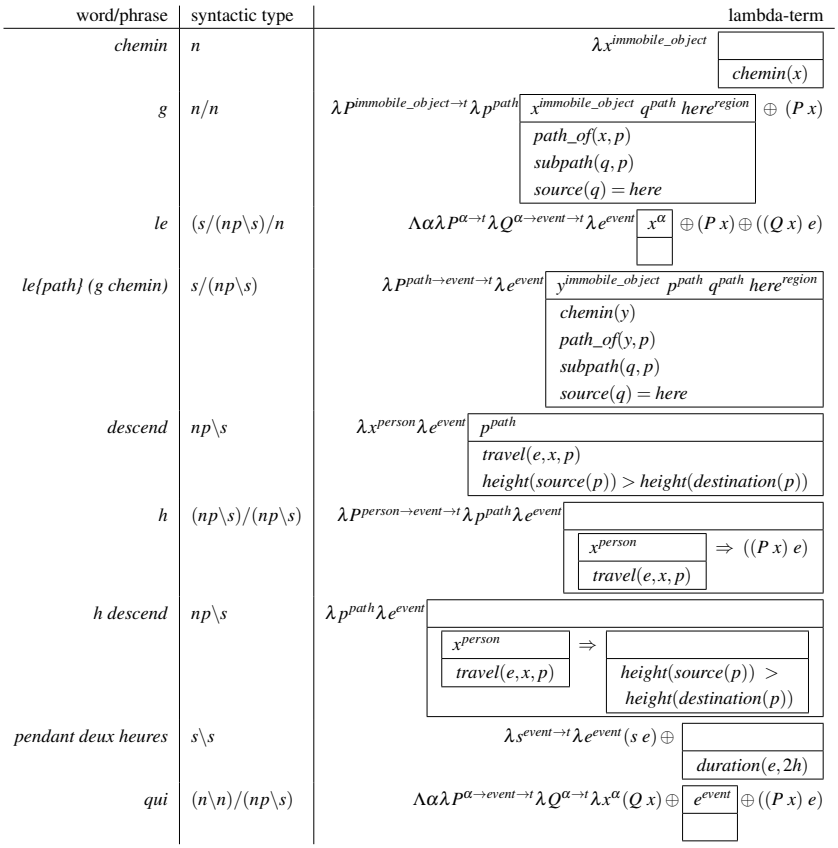
un groupe verbal requiert un sujet humain (*person*), ou en tout cas mobile, tandis que le dit sujet est une route, un chemin (*path*), etc. À la différence des phénomènes traités antérieurement, il n'est pas possible que le terme assurant le changement de type soit une simple constante de changement de type, puisqu'il doit contenir la variable correspondant au voyageur (liée par une quantification existentielle négative, insérée dans une conditionnelle ou par une quantification universelle).

Clairement, c'est u^{path} qui produit un x^{person}, mais si u restait argument de P devenu *person* on serait confronté aux deux problèmes suivants :

1. D'une part, le quantificateur correspondant au voyageur fictif ne pourrait avoir la portée sur le prédicat.
2. D'autre part les propriétés du chemin, comme par exemple, *goudronné*, deviendraient des propriétés du voyageur fictif !

Nous donnons en figure 3 les λ-termes de l'exemple (8). Une paraphrase possible serait : si un être humain suit ce chemin, alors il descend. Vu le type de ce terme, on peut lui appliquer des modifieurs comme *pendant deux heures* ou *pendant trois kilomètres* de la manière usuelle.

Figure 3- Types lexicaux et λ-termes pour “le chemin descend pendant deux heures”, avec les coercions de g et h.



Aux types *person*, *t*, *event*, *path*, on ajoute le type *region* afin de manipuler les données spatiales ainsi que *immobile_object*. Les prédicats nécessaires à la représentation sémantique des chemins sont *path_of* et *sub_path* le premier mettant en relation un objet immobile et l’entité de type *path* qui lui sera appliquée, et le second représentant la portion du chemin que l’on traite dans la phrase. *Source* et *destination* sont les fonctions permettant de manipuler les points départ et d’arrivée de la portion du chemin en question.

La coercion opérée par *h* permet de faire apparaitre le voyageur sous la forme x^{person} et de le faire entrer dans la relation *travel* avec une entité de type *path* (sans oublier l’entité de type *event*). La coercion *g* quant à elle permet de créer le lien entre l’objet immobile "chemin" et le type du chemin, *path* nécessaire à la construction d’un itinéraire.

Nous présentons ici deux exemples dont le traitement s’intègrent dans la même proposition.

(9) En effet, on est ici voisin de Toulouse ; comme le caractère, le type est nouveau. Les jeunes filles ont des figures fines, régulières, d’une coupe nette, d’une expression vive et gaie. Elles sont petites, elles ont la démarche légère, des yeux brillants, la prestesse d’un oiseau.

Ici, définir les filles de la région comme ayant des "figures régulières", étant "petites" ou encore ayant la "prestesse d’un oiseau" est une comparaison sous entendue au concept général recouvrant l’ensemble de toutes les filles. On conceptualise facilement une idée de la taille comme étant normale pour un spécimen du type "fille". Ainsi il nous faut un opérateur pouvant sélectionner toutes les propriétés telle que la taille d’un type particulier afin de pouvoir l’associer au spécimen de ce type, quelque soit le type concerné. En premier lieu, dans le système F nous rappelons qu’il n’existe qu’un quantificateur peu importe la classe d’objet sur

laquelle on quantifie, ce qui permet de quantifier sur tous les ordres. Au lieu d'avoir une constante $\forall_\alpha$ de type $(\alpha \to \mathbf{t}) \to \mathbf{t}$ pour chaque type α sur lesquels on voudrait quantifier, le quantificateur dans toute sa généralité est donc une constante $\forall$ de type $\Pi\alpha.(\alpha \to \mathbf{t}) \to \mathbf{t}$. Cette constante sera ensuite appliquée au type sur lequel on désire quantifier. Nous introduisons une constante $\measuredangle$ (Retoré, 2012) de type $\Pi\alpha.(\alpha \to \mathbf{t}) \to \alpha$ qui associe à une propriété P de type $\alpha \to \mathbf{t}$ (P est une propriété des objets de type α) l'élément générique de type α. La propriété P pour cet élément est vraie lorsque la plupart des éléments de type α ont la propriété P, on pense ici à l'analogie avec le $\tau x.A$ d'Hilbert (1922)[5].

Notre deuxième exemple concerne la prédication sur un ensemble d'individus dont on peut interpréter de deux manières divergentes la sémantique.

(10) Edgar et son guide descendaient toujours ensemble !... Enfin, le groupe allait se briser sur une saillie de roc effrayante, quand Vincent se précipita avec intrépidité au-devant d'eux, enfonçant par un coup désespéré sa hache tout entière dans la neige...

Cet exemple du corpus permet d'observer un phénomène courant en français où l'on peut recouvrir intuitivement le concept d'un ensemble d'individus agissant collectivement ou d'un ensemble d'individus agissant individuellement. Ainsi on peut comprendre dans "le groupe allait se briser sur une saillie de roc effrayante" que la chute a séparé les deux individus qui le composaient, dans ce cas c'est le groupe qui se brise ou encore que chacun d'entre eux a subit les dommages de l'accident, et alors ce sont les individus appartenant au groupe qui sont brisés. Suivant l'interprétation choisie, la valeur de vérité sera vrai pour l'un et faux pour l'autre.

Le système F offre la possibilité de gérer cette difficulté et les deux interprétations grâce à la constante de distributivité $* : \Lambda\alpha\lambda P^{\alpha\to t}\lambda Q^{\alpha\to t}\forall x^{\alpha}.Q(x) \Rightarrow P(x)$ qui permet une distributivité de la propriété sur les membres de l'ensemble. Les détails formels concernant cette constante et d'autres gérant la coercion et la distributivité stricte sont décrits dans Moot and Retoré (2011).

6 Conclusion

Dans cet article nous avons montré les différentes étapes de notre traitement automatique du discours, consistant en l'analyse syntaxique puis en la dérivation sémantique en λ-DRT. L'interface syntaxe-sémantique dans le cadre de la théorie des types est une base solide permettant de respecter la compositionnalité du sens tout en s'appuyant sur l'organisation syntaxique du discours. Le système F quant à lui est approprié pour traiter les phénomènes rencontrés dans le discours et l'interface sémantique-pragmatique justifie un raffinement du lexique par ce système. Pour de futurs travaux, nous envisageons d'enrichir davantage le lexique afin de couvrir plus largement le discours et les phénomènes de glissement de sens tels que nous les avons présentés. Plus spécifiquement, dans le cadre du projet Itipy il est nécessaire d'ordonner temporellement les évènements lorsqu'ils traitent du déplacement du voyageur dans le récit. Nous comptons développer le traitement de la temporalité des évènements dans ce même formalisme et dans l'esprit des travaux de Verkuyl (2008) tel qu'initié dans (Lefeuvre et al., itre).

Références bibliographiques

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In *Treebanks*. Kluwer.

Asher, N. (2011). *Lexical Meaning in context – a web of words*. Cambridge University Press.

Bassac, C., Mery, B., and Retoré, C. (2010). Towards a Type-Theoretical Account of Lexical Semantics. *Journal of Logic Language and Information*, 19(2) :229–245.

Busquets, J., Vieu, L., and Asher, N. (2001). La SDRT : Une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum*, XXIII(1) :73–101.

Ewald, W. B. (1996). *From Kant to Hilbert. A Source Book in the foundations of Mathematics*, volume 2. Oxford University Press, Oxford.

Girard, J.-Y. (1971). Une extension de l'interprétation de Gödel à l'analyse et son application : l'élimination des coupures dans l'analyse et la théorie des types. In Fenstad, J. E., editor, *Proceedings of the Second Scandinavian Logic Symposium*, volume 63 of *Studies in Logic and the Foundations of Mathematics*, pages 63–92, Amsterdam. North Holland.

Hilbert, D. (1922). Die logischen grundlagen der mathematik. *Mathematische Annalen*, 88 :151–165.

Howard, W. A. (1980). The formulae-as-types notion of construction. In Hindley, J. and Seldin, J., editors, *To H.B. Curry : Essays on Combinatory Logic, λ-calculus and Formalism*, pages 479–490. Academic Press.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

Kien, E. (2010). Du sens des mots à l'analyse automatique d'une phrase. Mémoire de stage d'initiation à la recherche, ENS-Cachan & INRIA Bordeaux.

Lefeuvre, A., Moot, R., Retoré, C., and Sandillon-Rezer, N.-F. (à paraître). Traitement automatique sur corpus de récits de voyages pyrénéens : Une analyse syntaxique, sémantique et temporelle. In *TALN'2012*.

Luo, Z. (2011). Contextual analysis of word meanings in type-theoretical semantics. In Pogodalla, S. and Prost, J.-P., editors, *LACL*, volume 6736 of *Lecture Notes in Computer Science*, pages 159–174. Springer.

Magri-Mourgues, V. (2009). *Le voyage à pas comptés. Pour une poétique du récit de voyage au XIXème siècle*. Number 9 in Lettres numériques. Honoré Champion.

Montague, R. (1974). *The Proper Treatment of Quantification in Ordinary English*, chapter 1. Blackwell Publishers.

Moot, R. (1998). Grail : An automated proof assistant for categorial grammar logics. In Backhouse, R., editor, *Proceedings of the 1998 User Interfaces for Theorem Provers Conference*, pages 120–129.

Moot, R. (2010). Wide-coverage French syntax and semantics using Grail. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

Moot, R. and Retoré, C. (2011). Second order lambda calculus for meaning assembly : on the logical syntax of plurals. In *Coconat 2011*.

Pasquali, A. (1994). *Le Tour des horizons : critique et récit de voyage*. Klincksieck.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Retoré, C. (2000). Systèmes déductifs et traitement des langues : un panorama des grammaires catégorielles. *Technique et Science Informatiques*, 20(3) :301–336. Numéro spécial *Traitement Automatique du Langage Naturel* sous la direction de D. Kayser et B. Levrat. Version préliminaire RR-3917 http ://www.inria.fr/.

Retoré, C. (2012). Variable types for meaning assembly : a logical syntax for generic noun phrases introduced by "most". *Recherches linguistiques de Vincennes*, 41 :1–18.

Verkuyl, H. J. (2008). *Binary Tense*. CSLI Publications.

Notes

[1]http ://www.lepoint.fr/actualites-societe/2010-05-24/ solidarite-le-lundi-de-pentecote-ne-fait-pas-recette/920/0/458325, visited 24 May 2010

[2]Le partage du type des entités n'est pas présenté dans les exemples de cet article.

[3]cette finitude n'est pas nécessaire, mais raisonnable aussi bien d'un point de vue applicatif que cognitif : les constantes sont introduites dans le lexique, qui comporte un nombre fini d'entrées, chacune ne contenant qu'un nombre fini de termes finis : par opposition aux modélisations en termes de mondes possibles, nous restons ici dans le champ des règles et du calculable.

[4]Comme nous prenons en compte les glissements de sens, et donc les changement de type de base, ce n'est pas aussi immédiat que dans la sémantique de Montague.

[5]pour une traduction partielle de l'article on optera pour (Ewald, 1996)